Survey on Sampling Based Intrusion-Detection System using Machine Learning Technique

Anurag Tripathi[1], Anurag Shrivastava[2]

[1]MTech Scholar, [2]Asso. Prof.

Department of CSE, NIRT, Bhopal

Abstract— Information security is becoming a more important issue for modern computer generation, progressively. Intrusion Detection System (IDS) as the main security defensive technique and is widely used against many category of attacks. Intrusion Detection Systems are used to detect various kinds of attack in very large datasets Data Mining and Machine Learning techniques has been proved useful and attract attention in the network intrusion detection research area. Recently, many machine learning methods have also been introduced by researchers, to obtain highly accurate and good detection rate. Unfortunately a potential drawback of all those methods is that how to classify abnormal activities or intrusion effectively. Machine learning algorithms is also used for obtain the high detection rate. Also, use of internet is increasing progressively, so that large amount of data is also an issue. Problem of large dataset can be solved using efficient sampling technique. This paper proposes a Hybrid sampling technique for obtaining the sampled and balanced dataset. Sampled dataset represent the whole dataset with accurate class balancing. This paper presents a Proposed Hybrid sapling and classifier architecture framework which is useful for improve classifier performance. The techniques were tested based on Detection rate and False Alarm rate. The result analysis and evaluation obtained by applying these approaches to the KDD CUP'99 data set.

Keywords— Intrusion Detection System (IDS), Classification, Machine Learning techniques, Hybrid sampling, KDD CUP'99 dataset.

I. INTRODUCTION

Secure information either in private or government sector has become an essential requirement. System vulnerabilities and valuable information catch the attention of most attackers'. Traditional intrusion detection approaches such as firewalls or encryption are not sufficient to prevent system from all attack types. The number of attacks through network and other medium has increased dramatically in recent years. Efficient intrusion detection is needed as a security cover against these malicious or suspicious and abnormal activities. Thus, classification in intrusion detection systems (IDSs) has been introduced as a security technique to classify & detect various attacks. The misuse detection and anomaly detection are two different techniques identified in IDSs. Misuse detection techniques can detect known attacks by examining attack patterns, similar to virus detection by an antivirus application. This technique require updated attack pattern to detect unknown attacks. On the other hand, anomaly detection identifies any unusual activity pattern which deviates from the normal usage, as intrusion. Although anomaly detection has the ability to detect unknown attacks which cannot be addressed by misuse detection, it suffers from high false alarm rate.

In recent years, and interest was given into machine learning techniques to overcome the constraint of traditional intrusion techniques by increasing accuracy and detection rates. Prediction and Description are two fundamental goals of Data Mining. Several data mining techniques like association, classification and clustering fulfill these goals. Classification and Machine learning techniques will be the main focus of this research. Classification is assigning a class label to a set of unclassified objects described by a fixed set of attributes or features. On the other hand, the field of machine learning is concerned with the higher-level question of how to construct computer programs that automatically learn by experience. In literature, numbers of IDS are developed based on many different machine learning techniques. Some researchers give the sampling technique for obtain the sampled data. The proposed techniques used sampled dada and developed as classifiers, which are used to classify or distinguish whether the incoming Internet access is the normal access or an attack.

II. MACHINE LEARNING TECHNIQUE

When a computer needs to perform a certain task, a programmer's solution is to write a computer program that performs the task. A computer program is a piece of code that instructs the computer which actions to take in order to perform the task. The area of machine learning is concerned with the higher-level question of how to make computer programs that automatically learn with experience. A computer program is said to learn from knowledge K with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with

experience E.[1] Thus, machine learning algorithms automatically extract knowledge from machine readable information. In machine learning, computer algorithms (learners) attempt to automatically collect knowledge from example data.

III. NEURAL NETWORKS

Artificial neural networks consists of a collection of processing elements that are highly interconnected and transform a set of inputs to a set of desired outputs that is inspired by biological nervous systems, such as the brain, process information. The technique of Neural Networks NN follows the same theories of how the human brains works.[11]

- DECISIONTREE

Decision tree is a classification technique of data mining for predictive models. Decision tree technique is like tree structure where internal node represents a test on attribute, branch represents an outcome of the test and leaf node represents a class label. From the pre classified data set it inductively learns to construct the models. Here each and every data item is defined by the attribute values. primarily decision tree is constructed by set of pre- classified data. The important approach is to select the attributes, which can best divide the data items into their respective classes based on these attributes the data item is partitioned [11].

- K-NEAREST NEIGHBOUR

K-nearest neighbour (k-nn) is a type of lazy learning, it simply stores a given training tuple and waits until it is given a test tuple. it is an instance based learner that classifies the objects based on closet training examples in the feature space. for a given unknown tuple, a k-nearest neighbour looks the pattern space for the k-training tuples that are closest to the unknown tuple. Comparison to other algorithm it is simplest machine learning algorithms.

- BAYESIAN CLASSIFIER

Bayesian Classifier algorithms provides high accuracy and speed for handling large dataset. In network model Bayesian classifier encodes the probabilistic relationship among the variable of interest. In intrusion detection this classifier is combined with statistical schemes to produce higher encoding interdependencies between the variables and predicting events. The graphical model of informal relationships performs learning technique. This technique is defined by two components-a directed acyclic graph and a set of conditional probability tables. Direct Acyclic Graph (DAG) represents a random variable, which may be discrete or continuous. For each variable classifier maintain one conditional probability table (CPT) and it requires higher computational effort.[11]

IV. RELATED WORK

Authors [1] conclude the method of IDS system for classification based method to rectify the dos, probe types attack. The various attacks contribute for making consider through various data set of using knowledge discovery set. So we will conclude in this paper for attacks and prevent them for using machine learning approaches.

In this work [2]. authors compared those methods and obtained the preliminary results using the sample data from Kyoto 2006+ data set. The RBF classification technique worked the best with an ROC value of 0.9741, whereas the Ensemble method came close with ROC of 0.9631. Although the Ensemble technique used in this study did not provide the best result, the technique has potential and further work in this area is worth-pursuing. While KDD-Cup 1999' data set has been used by more than 50% of the researchers working in the network security area [1], there is only limited work performed using Kyoto 2006+ data set.

In this work [6] authors use a data-driven approach based on an under- sampling technique to evaluate the performance of classifiers in the detection of network intrusion. We applied an under-sampling technique in two IDS datasets and evaluated the performance of 5 (five) classifiers in a 5% dataset portion followed by a validation step in a 2% portion of the same dataset without overlaps or repetitions. The results indicate that the use of a stratified train/test into under-sampled datasets show stability in the results in relation to the validation sub sampling. The use of this approach allows us to evaluate the classifiers in reduced time, including those considered computationally costly.

Hua TANG et al. [7] proposed a new approach to detect network attacks, which aims to study the efficiency of the method based on machine learning in intrusion detection, including artificial neural networks and support vector machine. The experimental results obtained by applying this approach to the KDD CUP'99 data set demonstrate that the proposed approach performs high performance, especially to U2R attacks and R2L type.

According to the authors [8], neural networks, SVM and decision trees are the popular schemes borrowed from Machine learning community into IDS. In it these three techniques are compared by applying on KDD CUP'99 data set. The machine learning approaches are supposed to be fit to identify the anomalies detection, in an appropriate way by proper training but the performace may be variable in terms of different algorithms.

In [9] authors said Intrusion detection system plays a major role in providing security to computer networks. The Intrusion detection system deals with large amount of data which contains various irrelevant and redundant features resulting in increased processing time and low detection rate. Therefore feature selection plays an important role in intrusion detection. There is various feature selection methods used. Authors compare the different feature selection methods are presented on KDDCUP'99 benchmark dataset and their performance are evaluated in terms of detection rate, root mean square error and computational time.

DRAWBACKS OF EXISTING TECHNIQUES

Several issues come from the survey such as large dataset, false detection, low detection rate, classification of attacks etc. To overcome the problem it is necessary to use sampling and learning technique for IDS and to test its results with existing technique. In this paper, a sampling and learning based technique is proposed which will reduce the dataset, improve detection rate with low false alarm rate. its results will be analyzed with existing technique

V. DATASET DESCRIPTION AND PRAPOSED HYBRID SAMPLING

5.1KDDCUP1999DATASET:

DARPA in concert with Lincoln Laboratory at MIT launched the DARPA 1998 dataset for evaluating IDS. The refined version of DARPA KDD Cup 1999 Data (KDD99) is used in the evaluate machine learning technique. Due to the computing power, we do not use the full dataset of KDD99 in the experiment but a 10% portion use of it. This 10% KDD99 dataset contains 494,021 records (each with 41 features) and 4 categories of attacks. The details of attack categories and specific types are shown in Table1. The four attack types are [2, 5, 7] :

(1) Probing: Scan networks to gather deeper information

(2) DoS:Denial of service

(3) U2R:Illegal access to gain super user privileges

(4) R2L:Illegal access from a remote machine

Every attack categories contain some specific attack types [2]. The number of specific attack types is different between 10% KDD99 dataset and full KDD99 dataset; we believe that there are no negative effects on our evaluation purpose.

Table1: Various types of Attack described in Four Major category

Attack Class	Attack Type
DoS	apache2, back, land, mailbomb, neptune , pod, processtable, smurf, teardrop,dpstrom
Probe	ipsweep,mscan,nmap, portsweep, saint
R2L	ftp_write,guess_passwd, imap, multihop, named, phf,sendmail,spy,snmpgetattack, snmpguess war ezclient, warez master, worm, xlock, xsnoop
U2R	buffer_over flow, httptunnel, loadmodule, perl, ps, rootkit, sqlattack, xtern

5.2 Sampling

Sampling is a technique employed for selecting a subset of the data. Sampling is the process of selecting representative which indicates the whole data set by examining a part. Sampling is needed in order to make abstraction of complex problem as well as it is used to acquire a sub set that is inferring a larger data set. It is widely accepted that a fairly modest-sized sample can sufficiently characterize a much larger population. [4] Sampling is used for the balanced, the highly imbalanced majority and minority classes. In KDD 99 data set having many imbalanced classes.

VI. PROPOSED CLASSIFIER FRAMEWORK

Firstly, we build the experiment evaluation environment with major steps: environment setup, data sampling is the first step Hybrid sampling software uses database management system and developer tools to develop the implementation as software setup with is software we obtain the sampled dataset. Now we start preprocessing step on Weka tool. Taking sampled data as input on Weka and apply supervised "CfsSubsetEval" filter for Feature selection, process. After applying Feature selection the number of features selected. The features selected are most relevant features with the Attack classification field and having less correlation among them. Now apply resample for class balancing. After apply resample all different class of dataset has balanced. Now in classification phase choosing the Decision tree classifier and split the training and testing data on the 60:40 Ratio.

Finally, come up with the performance comparison between the sampled data and original data on the setup environment.

In order to verify the effectiveness of different

classifiers for the field of intrusion detection, we will use the sampled KDD99 dataset to make relevant experiments step-by-step.

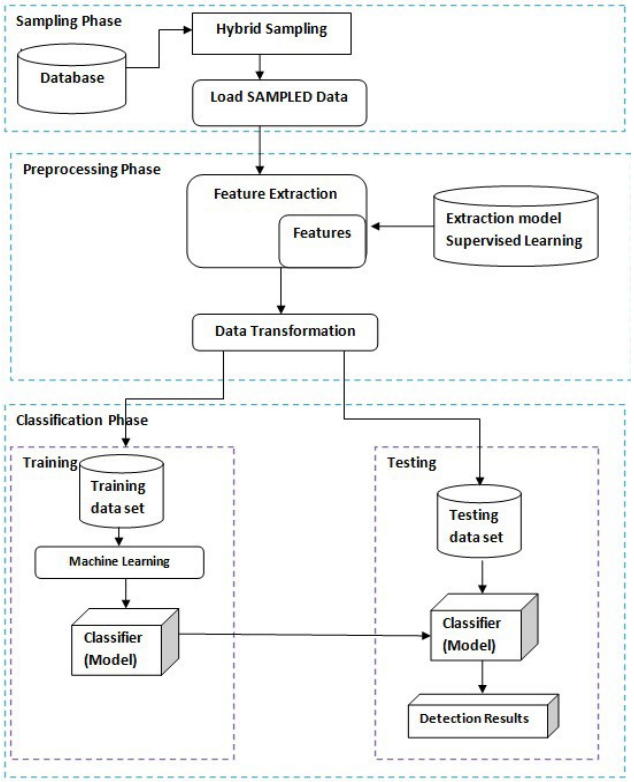

Fig.1PROPOSEDFRAMEWORK

Following fundamental definition and formulas are used to estimate the performance of the classifier: accuracy rate (AR) and Error Rate (ER).

True Positive: When, the number of found instances for attacks is actually attacks.

False Positive: When, the number of found instances for attacks is normal.

True Negative: When, the number of found instances is normal data and it is actually normal.

False Negative: When, the number of found instances is detected as normal data but it is actually attack.

The accuracy of IDS classifier is measured generally on basis of following parameters:

Detection Rate: Detection rate refers to the percentage of detected Attack among all attack data, and is defined as follows:

$$\text{Detection rate} = \frac{TP}{TP+TN} * 100$$

With this formula detection rate for different types of Attacks can be calculated.

False Alarm rate: False alarm rate refers to the percentage of normal data which is wrongly recognized as attack. , and is defined as follows:

$$\text{False Alarm rate} = \frac{FP}{FP+TN} * 100$$

With this formula false alarm rate for different types of Attacks can be calculated.

VII. CONCLUSION

Intrusion Detection System (IDS) as the main security defensive technique in Information security. In recent years Machine Learning techniques proved useful and attracted increasing attention in the network intrusion detection research area. Looking at the need of classification of correct attacks. This paper concluded by suggesting a framework for Classification and method to evaluate the framework. Sampling has been often suggested as an efficient tool to reduce the size of the dataset operated at some cost to accuracy. But, proposed framework with hybrid sampling in this work is showing greater accuracy when tested with three classifiers. The proposed framework having Hybrid Sampling is effective and extensible in terms of various performance parameters. Hybrid sampling based intrusion detection system gives better accuracy and detection rate.

The future works focus on same formwork tested with more machine learning classifiers, and implement IDS for actual detection of attack in real system. Sampling technique can be further compared with other sampling technique. Combination of other sampling with proposed algorithms also suggested.

REFERENCES

[1] Doodipalli Subramanyam, "Classification of Intrusion Detection Dataset using machine learning Approaches" 978-1-5386-7709- 4/18/$31.00 c 2018IEEE

[2] Marzia Zaman, Chung-Horng Lung "Evaluation of Machine Learning Techniques for Network Intrusion Detection" 978-1- 5386-3416-5/18/$31.00 ©2018 IEEE

[3] Anushka Srivastava, Avishka Agarwal, Gagandeep Kaur "Novel Machine Learning Technique for Intrusion Detection in Recent Network-based Attacks " 978-1-7281-3651-6/19/$31.00 ©2019 IEEE

[4] Kazi Abu Taher, Billal Mohammed Yasin Jisan, Md. Mahbubur Rahman "Network Intrusion Detection using Supervised Machine Learning Technique with Feature Selection" 978-1- 5386-8014-8/19/$31.00 ©2019 IEEE

[5] Anish Halimaa A, Dr. K Sundarakantham "MACHINE LEARNING BASED INTRUSION DETECTION SYSTEM "978-1-5386-9439-8/19/$31.00 ©2019 IEEE

[6] Bruno Silva, Manuel Silva Neto, Paulo Cortez and Danielo Gomes "Design of Network Intrusion Detection Systems with under-sampled datasets" 978-1-7281-3185-6/19/$31.00 c 2019 IEEE

[7] Hua TANG, Zhuolin CAO “Machine Learning-based

Intrusion Detection Algorithms" Journal of Computational Information Systems5:6(2009) 1825-1831 Available at http://www.JofCI.org

[8] YU-XIN MENG "The Practice on Using Machine Learning For Network Anomaly Intrusion Detection" 2011 IEEE.

[9] Megha Aggarwal, Amrita "Performance Analysis Of Different Feature Selection Methods In Intrusion Detection" international journal of scientific & technology research volume 2, issue 6, june 2013.

[10] Devendra kailashiya Dr. R.C. Jain Improve Intrusion Detection Using Decision Tree with Sampling" IJCTA | MAY-JUNE 2012

[11] Kamarularifin Abd Jalill, Mohamad Noorman Masrek "Comparison of Machine Learning Algorithms Performance in Detecting Network Intrusion" 201O International Conference on Networking and Information Technology 978-1-4244-7578- 0/$26.00 © 2010 IEEE

[12] Yunming Ye a,e,n, QingyaoWua,e, JoshuaZhexueHuang b,d, MichaelK.Ng c, XutaoLi "Stratified sampling for feature subspace selection in random forests for high dimensional data" Contents lists available at SciVerse Science Direct journal homepage: www.elsevier.com/locate Pattern Recognition September 2012.

[13] Ligang Zhou "Performance of corporate bankruptcy prediction models on imbalanced dataset:The effect of sampling methods Contents lists available at SciVerse ScienceDirect 2013

[14] Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, and Ali A. Ghorbani "A Detailed Analysis of the KDD CUP 99 Data Set" Proceedings of the 2009 IEEE Symposium on Computational Intelligence in Security and Defense Applications (CISDA 2009)

[15] Megha Aggarwal, Amrita "Performance Analysis Of Different Feature Selection Methods In Intrusion Detection" international journal of scientific & technology research volume 2, issue 6, june 2013

[16] Huy Anh Nguyen and Deokjai Choi "Application of Data Mining to Network Intrusion Detection: Classifier Selection Model"

[17] Arun K Pujari "Data mining techniques" Universities Press.

[18] Tahir Mehmood and Helmi B Md Rais "Machine learning Algorithms in context of Intrusion Detection" 978-1-5090-2549- 7/16/$31.00 ©2016 IEEE 978-1-5090-2549-7/16/$31.00